Landmark based navigation in next generation Vehicles

Mohan Krishna Manchala

A Thesis Submitted to

Indian Institute of Technology Hyderabad

In Partial Fulfillment of the Requirements for

The Degree of Master of Technology

Department of Electrical Engineering

July 2014

Declaration

I declare that this written submission represents my ideas in my own words, and where ideas or words of others have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the Institute and can also evoke penal action from the sources that have thus not been properly cited, or from whom proper permission has not been taken when needed.

(Signature)

M. Mohan Krishna

(Mohan Krishna Manchala)

CONBM01

(Roll No.)

Approval Sheet

This thesis entitled – *Landmark based navigation in Next Generation vehicles* – by – *M. Mohan Krishna* – is approved for the degree of Master of Technology from IIT Hyderabad.

Dr. Phanindra Jampana
Dept. of Chemical Engineering
Examiner

Dr. Sumohana Channappayya
Dept. of Electrical Engineering
Examiner

Dr. Soumya Jana
Dept. of Electrical Engineering
Advisor

Dr. C. S. Sastry
Dept. of Mathematics
Chairman

Acknowledgements

During my study at IIT Hyderabad, I have met many people who have provided priceless advice, support and care to me and my work. Without them, I would not have been able to finish my dual degree study.

First of all, I would like to thank my advisor, Dr. Soumya Jana, for his great guidance and numerous support to my research and life. I have learned from Dr. Soumya Jana not only how to conduct research, how to solve real world problems, how to collaborate with other researchers, etc. He always believed in me and provided opportunities beyond my capabilities. I am really glad to have the opportunity to work with Dr. Soumya Jana during my masters.

I would also like to thank my other thesis committee member, Dr. Sumohana Channappayya. Dr. Sumohana has encouraged me to convert from B.Tech to M.Tech dual degree and gave me the motivation to pursue higher studies. I have enjoyed working with him, and benefited greatly from the interactions with him.

I spent five years at IIT Hyderabad. I want to thank the following friends, who made my life here a pleasant and memorable experience: Sandeep (B.Tech), Sai Prasad(B.Tech), Ashok(B.Tech), Ashok Anand(B.Tech), Harsha(B.Tech), Roopak, Kiran, Sandeep, Harsha, Naresh and Anand. My apologies if I left anybody out. I also want to thank you if you are reading this dissertation.

Finally, I sincerely thank my family. My parents who always supported me no matter joy or hardship, success or failure. My brothers who always encouraged me. My relatives and other friends also deserve my thanks.

Dedication

To my family and friends.

Abstract

Most existing Global Positioning System (GPS)-based vehicle navigation systems (also termed route guidance systems) utilize distance within their turn-by-turn navigation directions. For example, a system might give a voice instruction such as "turn left in 0.2 mile". However, human drivers usually use landmarks to help their navigation. Some previous research has shown that there are performance-related benefits in using landmarks instead of distance for navigation. The goal of this thesis work is to develop a landmark-based navigation system in next generation cars using computer vision techniques. We also provide an interactive platform based on hand gestures to intuitively interact with the landmark-based navigation system.

Contents

Chapter 1

Introduction

For the past hundred years, innovation within the automotive sector has brought major technological advances, leading to safer, cleaner, and more affordable vehicles. But for the most part, the changes have been incremental, evolutionary. In this technology world, any innovation is only as good as the next thing coming down the line. Therefore now, in the early decades of the 21st century, the industry appears to be on the cusp of revolutionary changewith potential to dramatically reshape not just the competitive landscape but also the way we interact with vehicles and, indeed, the future design of our human transportation system.

Though we are familiar with the topic of military drones, next generation ideas such as driver-less or fuel-less cars sounds pretty far-fetched. But are they still just science fiction? Something that gets batted around in research labs and think tanks? Or are next generation vehicles on the verge of becoming a viable form of personal mobility? Will the market accept them, want them, and pay for them?

We think the answer is a resounding yes. The marketplace will not merely accept next generation vehicles; it will be the engine pulling the industry forward. Consumers are eager for new mobility alternatives that would allow them to stay connected and recapture the time and psychic energy they squander in traffic jams and defensive driving.

Researchers have been working on many technologies of next generation cars whether it be for safety, entertainment, usefulness or simply for pure innovation. The following are few of the technologies for next generation vehicles:

- Automatic navigation system
- Vehicle to vehicle interaction
- Energy producing body panels
- Augmented reality wind shield
- Collision mitigation system
- Gesture and Voice recognition
- Intelligent parking guidance
- Health monitoring

In this thesis work we focused on developing a landmark based navigation system and hand gestures recognition system for human vehicle interface.

1.1 Motivation

Navigation is the process of planning, recording, and controlling the movement of a craft or vehicle from one place to another. Navigating a vehicle in a dynamic environment is one of the most demanding activities for drivers in their daily lives. Americans drive 12,000 miles per year on average. Studies have long identified the difficulties that drivers have in planning and following efficient routes [1].

A vehicle navigation system (also termed route guidance system) is usually a satellite navigation system designed for use in vehicles. Most systems typically use a combination of Global Positioning System (GPS) and digital map matching to calculate a variety of routes to a specified destination such as the shortest route. They then present a map overview and turn-by-turn instructions to drivers, using a combination of auditory and visual information. A typical turn-by-turn instruction is an auditory prompt such as "in 0.5 mile turn right", accompanied by a visual right turn arrow plus a distance-to-turn countdown that reduces to zero as the turn is approached. Vehicle navigation systems generally function well, although they are wholly dependent on the accuracy of the underlying map database and availability of GPS signals. However, from a human factor perspective, there are several potential limitations to the current design [2]: mainly presenting procedural and paced navigation information to the driver, and relying on distance information to enable a driver to locate a turn.

Human drivers often use landmarks for navigation. For example, we tell people to turn left after the second traffic light and to make a right at Apollo hospital. In our daily lives, a landmark can be anything that is easily recognizable and used for giving navigation directions, such as a sign or a building.

It has been proposed that current navigation systems can be made more effective and safer by incorporating landmarks as key navigation cues [3]. Especially, landmarks support navigation in unfamiliar environments. By providing external reference points, which are easily remembered and recognized, landmarks can potentially reduce the need to refer to an information display in order to locate a navigation decision point.

The definition of landmark in navigation context has been studied from varying theoretical perspectives. Lynch described landmarks as external reference points that are easily observable from a distance[4]. Kaplan defined a landmark as a known place for which the individual has a well formed representation, and described two theoretical factors that lead to an object or place acquiring landmark status: the frequency of contact with the object or place, and its distinctiveness [5]. Three types of distinctiveness were proposed: visual distinctiveness (a predominantly objective quality relating to the physical attributes that discriminate a landmark from the surrounding environment); inferred distinctiveness (knowledge concerning its structure or form that makes the landmark stand out from what is usual); functional distinctiveness (the salience in terms of the goals or sub-goals of the landmark). In addition to the visual characteristics of landmarks and their functional or social importance, the location of an object within the environment has a significant impact on its effectiveness as a landmark. The following are the attributes that are most important characteristics of good landmarks for vehicle navigation systems:

1. Usefulness of location: the ease with which the location of the landmark allows a navigational manoeuvre (e.g., a turning) to be identified.

2. Visibility: whether the landmarks size and shape can be clearly seen in all conditions.

3. Uniqueness: whether the appearance of the landmark is such that it is unlikely to be mistaken for anything else.

4. Permanence: the likelihood of the landmark being present.

In our daily life, we observe that common navigation-useful landmarks include 1) road signs, 2) other signs (including signs of petrol bunks, restaurants, shops, etc.) and 3) buildings. This is our observation. In this thesis work we mainly focused on identifying landmark buildings.

(a) Road sign (b) Restaurant (c) Building

Figure 1.1: Common landmarks useful for navigation

1.2 Problem statement

In this thesis, we aim to develop technologies for landmark based navigation for the situations similar to the following scenario.

Ram and his friend Shyam plan to attend a party in a city not familiar to Ram, and they are driving separately. Both their cars have cameras that capture videos of the scenes along the route. Shyam records the video and identifies landmarks along the route using landmark identification technique and sends them to Ram. Ram uses these landmarks to match in his video sequence to validate his navigation. Also Ram would like to have a natural and intuitive interaction with his navigation system.

We would like to build such a system that could help Ram interactively navigate to the destination.

To implement such a landmark-based navigation system in next generation vehicles, I focus my thesis around the following two main problems:

1. Landmark based navigation system (Chapter 2): To identify good landmarks from the video sequence and matching the identified landmarks with a video to validate the route.

2. Hand Gesture recognition system (Chapter 3): To develop an intuitive interaction with the navigation system using hand gestures.

We aim to develop a landmark based navigation system in next generation vehicle. We mainly focus on identifying building landmarks. We present a saliency based landmark identification technique in images and videos. In addition, we also propose to integrate hand gesture recognition system to intuitively interact with the landmark based navigation system.

1.3 Related work

In this section we review some previous research on building landmark-based navigation.

The recognition of landmark buildings generally comprises two phases: building representation and recognition of the target building. The problem of building recognition has attracted much attention in the past, mostly considering outdoors scenes. Some researchers formulate and tackle the problem in a content-based image retrieval manner [6]. Other researchers proposed using vanishing direction for alignment of a building view in the query image to the canonical view in the database and proposed matching using interest regions descriptors, followed by the relative pose recovery between the views from planar homographies [7]. As mentioned in the paper, the methods which employ solely geometric and local feature based matching techniques are often slow. In [8] authors proposed extracting invariant regions and used a set of color moment invariants to represent them. Recognition was performed based on the number of matched regions. In [9], an alternative approach was proposed to the context-based place recognition problem. The representation of individual locations was obtained by integrating responses of the bank of filters over coarse spatial regions and fitting a Gaussian mixture model to the responses. This approach enabled coarse classification of locations and also exploited spatial relationships between locations captured by a Hidden Markov Model. However, the location model did not allow for actual pose recovery of the camera with respect to the scene.

In the context of object recognition, both global and local image descriptors have been considered. Commonly used global descriptors, which provide some invariance to occlusions and clutter proposed in the past, include gist features and multi-dimensional histograms. The representatives of local image descriptors include scale invariant features and their descriptors, which are invariant with respect to rotation, scale change and affine transformations. From the perspective of the application, the efficiency of the approach has to be considered. Therefore, when dealing with large databases, it is desirable to have some simple indexing vectors for all models, so that unlikely models can be eliminated in advance. More recent research on detection and recognition of buildings has been reported in [10].

1.4 System architecture

The following is the system architecture of our landmark-navigation system in next generation vehicles (see Figure 1.2).

The system comprises of two main functions. 1) Landmark-based navigation and 2) Hand gesture recognition.The Landmark-based navigation system handles the reference and guidance modules. Reference module identifies possible landmarks along the path to the destination and guidance module matches the identified landmarks along the path and guides the navigation. The hand gesture recognition system helps the user to manually validate the reference and guidance modules

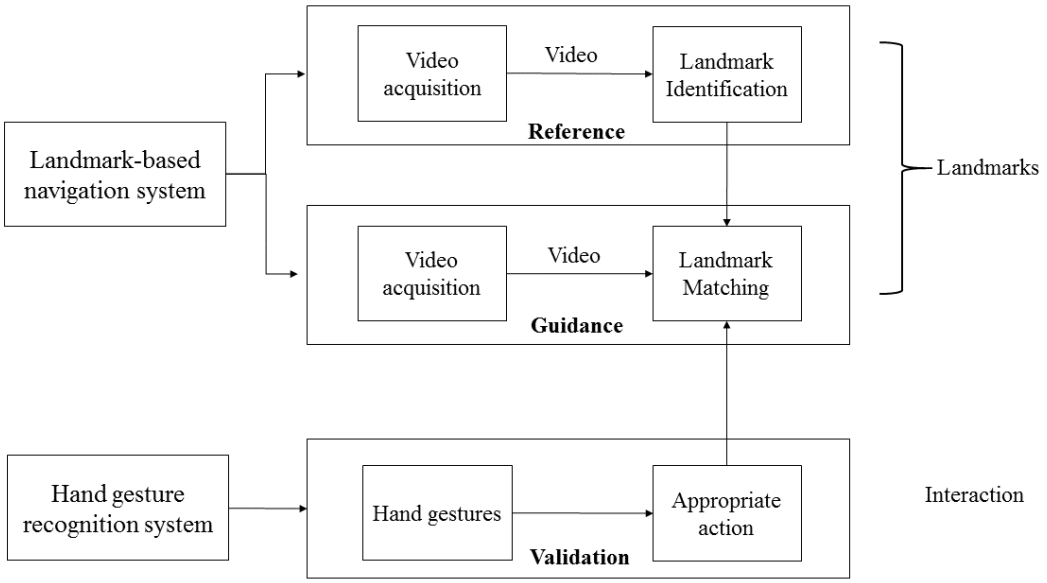


Figure 1.2: System architecture

through natural hand gestures. It gives the user a platform to intuitively interact with the navigation system.

1.5 Organization of the thesis

In the Chapter 2, we will discuss our proposed landmark-based navigation system. In the Chapter 3, we will discuss a hand gesture recognition system which provides a user a platform to interact with the landmark-based navigation system. Finally, we will conclude this thesis and present some future works in the Chapter 4.

Chapter 2

Landmark based navigation system

Landmark buildings are an important class of street landmarks which can be used as reference points for driving navigation. Accurately recognizing landmark buildings poses many challenges. Most buildings are occluded by nearby objects such as trees. Hence, different view angles may result in different occlusion appearance. Appearance changes due to lighting, environmental changes, weather or season changes. These challenges prevent us from applying direct object ngerprint-based approach to recognize buildings. The navigation system problem as considered in this thesis consists of two phases: Landmark identification and Landmark matching. Landmark identification has been done by segmenting the landmarks based on region contrast saliency object segmentation technique. The database of the identified landmarks is used to guide the navigation of the user by matching them along the destination path. SURF features were utilized for landmark matching. In this work, we only focus on the recognition aspect within the driving context. Iterate our application context as mentioned in the problem statement. Ram and his friend Shyam plan to attend a party in a city not familiar to Ram, and they are driving separately. Both their cars have cameras that capture videos of the scenes along the route. Shyam records the video and identies landmarks along the route using landmark identification technique and sends them to Ram. Ram uses these landmarks to match in his video sequence to validate his navigation. In this chapter, we would like to build a system that could help Ram automatically navigate to the destination by matching the identified target landmark buildings along the path.

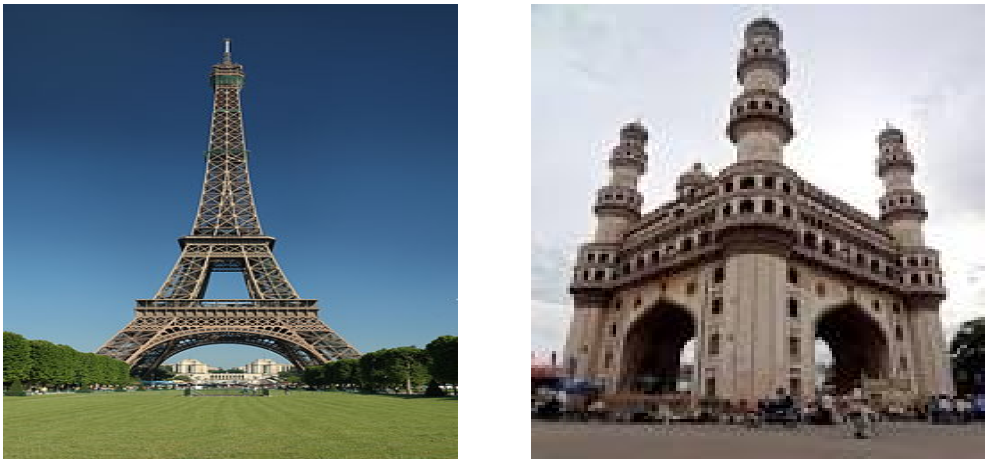
2.1 Landmark Identification

A landmark is a recognizable natural or man-made feature used for navigation, a feature that stands out from its near environment and is often visible from long distances. Originally, a landmark literally meant a geographic feature used by explorers and others to find their way back or through an area. For example the Table Mountain near Cape Town, South Africa, is used as the landmark to help sailors to navigate around southern tip of Africa during the Age of Exploration. Other than natural geographic feature, man-made structures are sometimes built to assist sailors in naval navigation. The Lighthouse of Alexandria and Colossus of Rhodes for example are ancient structures from antiquities built for this purpose, to lead ships to the port. Humans often tend to associate a landmark to signify a particular place. For instance, Hyderabad city is associated with Charminar,

Agra is symbolized with Taj Mahal, Newyork city is colligated with statue of liberty and Cape Town, South Africa is signified by Table Mountain.

(a) (b)

Figure 2.1: Natural landmarks (a) Table mountain and the ocean at cape town, and (b) Niagara falls



(a) (b)

Figure 2.2: Man-made landmarks (a) Eiffel Tower, and (b) Charminar

Human drivers remember landmarks for a better navigation. This helps us easily identify the route to the destination without any confusion. A good landmark would be more beneficial for an optimized navigation. Landmarks which are distinctive from their surroundings, which have wide visibility and which provide many visual cues are often considered as good landmarks. Landmarks are usually classified as either natural landmarks or man-made landmarks as shown in the figure, both can be used to support navigation on finding directions. Natural landmarks consists of characteristic features, such as mountains or plateaus whereas man-made landmarks are usually referred to as monuments or distinctive buildings.

In this thesis, we have tried to identify only man-made landmarks. We propose a saliency based landmark identification and used SURF features for landmark matching.

2.1.1 Segmenting salient objects from images and videos

A good landmark is always distinctive from its surroundings. It is attributed to the variations in image features like color, gradients, edges, and boundaries compared to its surroundings. Salient object is an object which attracts human visual attention. Saliency originates from visual uniqueness, unpredictability, rarity, or surprise. Therefore we used saliency property of the landmark to segment it from the video frames.

Biological vision systems are remarkably effective in finding relevant targets from a scene[12] . Identifying these prominent, or salient, areas in the visual field enables one to allocate the limited perceptual resources in an efficient way. Compared to biological systems, computer vision methods are far behind in the ability of saliency detection. However, reliable saliency detection methods would be useful in many applications like adaptive compression and scaling [12], unsupervised image segmentation [13], and object recognition.

Early work by Treisman and Gelade [14], Koch and Ullman [15], and subse -quent attention theories proposed by Itti, Wolfe and others, suggest two stages of visual attention: fast, pre-attentive, bottom-up, data driven saliency extraction; and slower, task dependent, top-down, goal driven saliency extraction. Here we used Global Contrast based Salient Region Detection proposed by Cheng and other [16] to segment the landmarks from the video frames. We focus on bottom-up data driven saliency detection using image contrast. It is widely believed that human cortical cells may be hard wired to preferentially respond to high contrast stimulus in their receptive fields [23]. We propose contrast analysis for extracting high-resolution, full-field saliency maps based on the following observations:

- A global contrast based method, which separates a large-scale object from its surroundings, is preferred over local contrast based methods producing high saliency values at or near object edges.
- Global considerations enable assignment of comparable saliency values to similar image regions, and can uniformly highlight entire objects.
- Saliency of a region depends mainly on its contrast to the nearby regions, while contrasts to distant regions are less significant.
- Saliency maps should be fast and easy to generate to allow processing of large image collections, and facilitate efficient image classification and retrieval.

Histogram-based contrast method (HC) to measure saliency assign pixel-wise saliency values based simply on color separation from all other image pixels to produce full resolution saliency maps. We use a histogram-based approach for efficient processing, while employing a smoothing procedure to control quantization artifacts. As an improvement over HC-maps, we incorporate spatial relations to produce region-based contrast (RC) maps where we first segment the input image into regions, and then assign saliency values to them. The saliency value of a region is now calculated using a

global contrast score, measured by the regions contrast and spatial distances to other regions in the image.

Histogram Based Contrast

Based on the observation from biological vision that the vision system is sensitive to contrast in visual signal, we propose a histogram-based contrast (HC) method to define saliency values for image pixels using color statistics of the input image. Specifically, the saliency of a pixel is defined using its color contrast to all other pixels in the image, i.e., the saliency value of a pixel I_k in image I is defined as,

$$S(I_k) = \sum_{\forall I_k \in I} D(I_k, I_i), \tag{2.1}$$

where $D(I_k, I_i)$ is the color distance metric between pixels I_k and I_i in the $L^*a^*b^*$ space (see also [32]). Equation (2.1) can be expanded by pixel order to have the following form,

$$S(I_k) = D(I_k, I_1) + D(I_k, I_2) + ... + D(I_k, I_N), \tag{2.2}$$

where N is the number of pixels in image I. It is easy to see that pixels with the same color value have the same saliency value under this definition, since the measure is oblivious to spatial relations. Hence, rearranging (2.2) such that the terms with the same color value c_j are grouped together, we get saliency value for each color as,

$$S(I_k) = S(c_l) = \sum_{j=1}^{n} f_j D(c_l, c_j), \tag{2.3}$$

where c_l is the color value of pixel I_k, n is the number of distinct pixel colors, and f_j is the probability of pixel color c_j in image I . Note that in order to prevent salient region color statistics from being corrupted by similar colors from other regions, one can develop a similar scheme using varying window masks. However, given the strict efficiency requirement, we take the simple global approach.

Region Based Contrast

Humans pay more attention to those image regions that contrast strongly with their surroundings. Besides contrast, spatial relationships play an important role in human attention. High contrast to its surrounding regions is usually stronger evidence for saliency of a region than high contrast to far-away regions. Since directly introducing spatial relationships when computing pixel-level contrast is computationally expensive, we introduce a contrast analysis method, region contrast (RC), so as to integrate spatial relationships into region-level contrast computation. In RC, we first segment the input image into regions, then compute color contrast at the region level, and define the saliency for each region as the weighted sum of the region's contrasts to all other regions in the image. The weights are set according to the spatial distances with farther regions being assigned smaller weights.

Region contrast by sparse histogram comparison: Segmenting the input image into regions is done (see figure 2.3) using a graph-based image segmentation method which is reproduced

from [17]. Then the color histogram is built for each region as in HC method. For a region r_k, corresponding saliency value is computed by measuring its color contrast to all other regions in the image(see (2.4)),

Figure 2.3: Region segmentation reproduced from [17].

$$S(r_k) = \sum_{r_k \neq r_i} D_r(r_k, r_i), \tag{2.4}$$

where $w(r_i)$ is the weight of region r_i and $D_r(r_k, r_i)$ is the color distance metric between the two regions. Here we use the number of pixels in r_i as $w(r_i)$ to emphasize color contrast to bigger regions. The color distance (2.5) between two regions r_1 and r_2 is defined as,

$$D_r(r_1, r_2) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} f(c_{1,i}) f(c_{2,j}) D(c_{l,i}, c_{2,j}), \tag{2.5}$$

where$f(c_k, i)$ is the probability of the i-th color $c_{k,I}$ among all n_k colors in the k-th region r_k, $k = \{1, 2\}$. Note that we use the probability of a color in the probability density function (i.e. normalized color histogram) of the region as the weight for this color to emphasize more the color differences between dominant colors. Storing and calculating the regular matrix format histogram for each region is inefficient since each region typically contains a small number of colors in the color histogram of the whole image. Instead, we use a sparse histogram representation for efficient storage and computation.

Spatially weighted region contrast: We further incorporate spatial information by introducing a spatial weighting term in (2.6) to increase the effects of closer regions and decrease the effects of farther regions. Specifically, for any region r_k, the spatially weighted region contrast based saliency is defined as:

$$S(r_k) = \sum_{r_k \neq r_i} exp(-D_s(r_k, r_i)/\sigma_s^2) w(r_i) D_r(r_k, r_i), \tag{2.6}$$

where $D_s(r_k, r_i)$ is the spatial distance between regions r_k and r_i, and σ_s controls the strength of spatial weighting. Larger values of σ_s reduce the effect of spatial weighting so that contrast to farther regions would contribute more to the saliency of the current region. The spatial distance between

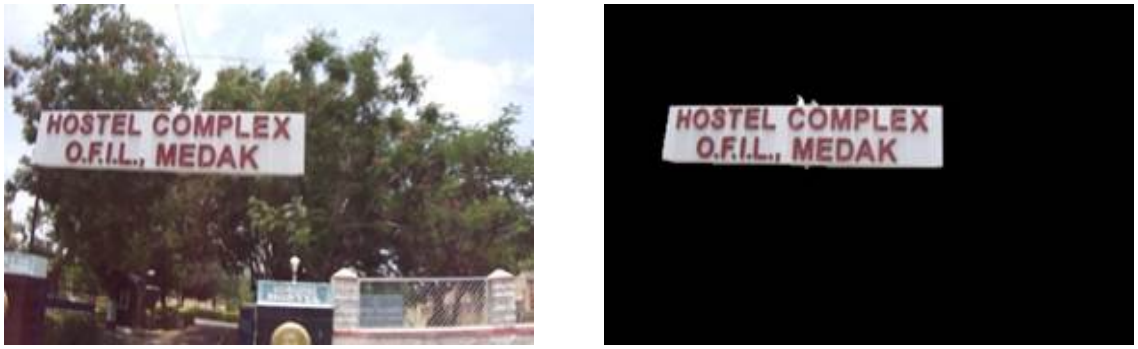
two regions is defined as the Euclidean distance between their centroids. In our implementation, we use $\sigma_s^2 = 0.4$ with pixel coordinates normalized to [0, 1].

Saliency cut: We now consider the use of the computed saliency map to assist in salient object segmentation. Saliency maps have been previously employed for unsupervised object segmentation. In our approach, we iteratively apply GrabCut [18] to refine the segmentation result initially obtained by thresholding the saliency map. Instead of manually selecting a rectangular region to initialize the process, as in classical GrabCut, we automatically initialize GrabCut using a segmentation obtained by binarizing the saliency map using a fixed threshold; the threshold is chosen empirically to be the threshold that gives 95% recall rate in our fixed thresholding experiments.

(a) A landmark building (b) Segmented landmark

Figure 2.4: Landmark identification

After doing saliency cut, we obtain saliency based landmarks as shown in the figures 2.4 and 2.5. Some landmark objects such as the water tank in the figure were not segmented due to their lack of salient texture when compared to the rest of the frame. In few frames we obtain the segmented regions of unwanted objects such as tress, sky, vehicles, etc as shown in the figure 2.6.



(a) Landmark: name of the building (b) Segmented landmark

Figure 2.5: Landmark identification

(a) Water tank landmark

(b) Unwanted object segmentation

Figure 2.6: Failed landmark identification

This happens when there are no salient landmarks in the frames and our method considers tress, sky, vehicles, etc. as salient regions. We remove such frames by employing an edge based unwanted object removal technique detailed in the next section.

2.1.2 Unwanted object removal

Many frames, after processing them through saliency based object segmentation contain unwanted objects which are not landmarks such as tress, sky, vehicles, etc. This will be due to occlusions and lack of salient texture for landmark buildings. In order to deal with such issues, we propose a edge information-based unwanted object removal technique. The edge detected image of a landmark building usually contains regular structures such as rectangles, straight lines and squares(see figure 2.7). Whereas the edge detected images of non-landmarks contain irregular structures which have more crooked lines and curves(see figure 2.8). Based on this logic we have devised an edge map based method(see figure 2.9) to discard the non-landmark frames.

Figure 2.7: Edge image of landmark building

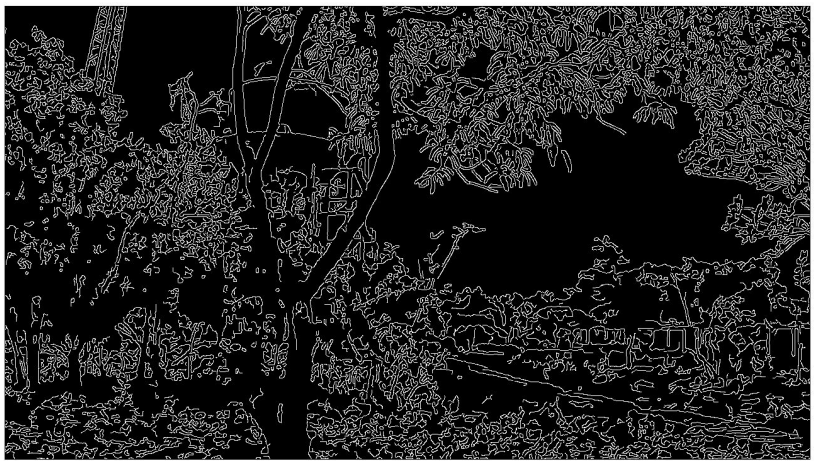


Figure 2.8: Edge image of trees

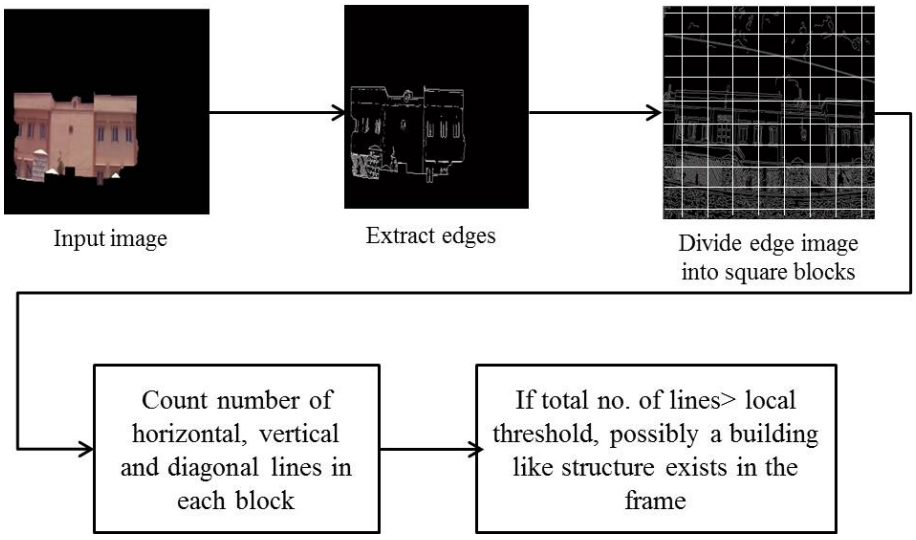


Figure 2.9: Edge map based unwanted object removal

Canny edge detection: Firstly we find the edge image of the frame using canny edge detector [19]. The edge detection process serves to simplify the analysis of images by drastically reducing the amount of data to be processed, while at the same time preserving useful structural information about object boundaries. We divide the edge detected image into square blocks of equal size i.e., 10x10 for further processing. For each block, we find the number of horizontal, vertical and cross straight lines and ignore the lines which have varying slope values. We do this for the entire image and calculate the total number of such lines.

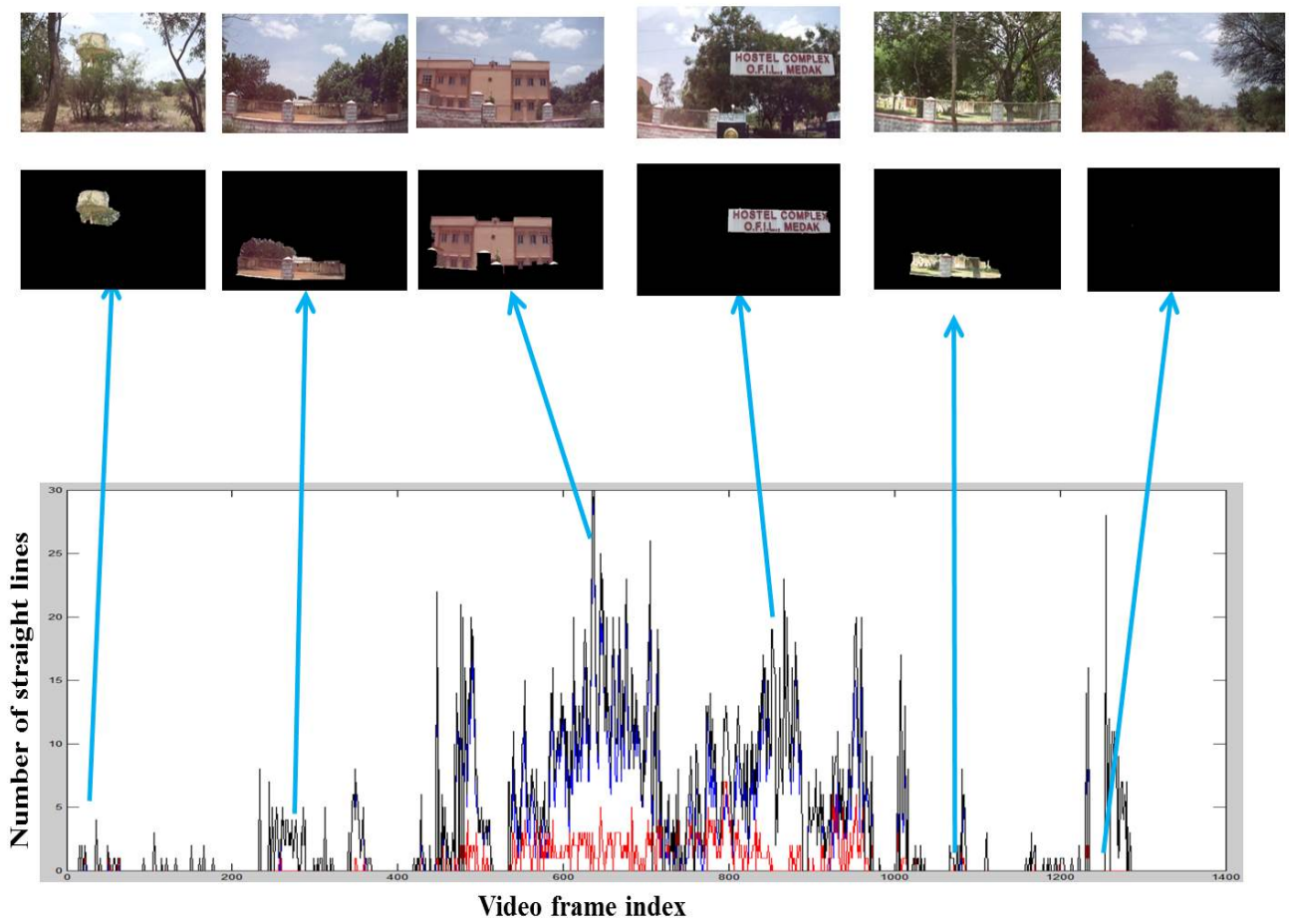


Figure 2.10

Now an edge map is constructed with the frame index along the x-axis and number of lines in the y-axis as shown in figure. We calculate the local maximum of number of straight lines for every 300 frames i.e., 10 seconds of the travel area and retain the corresponding frame. We discard all the other frames. By doing this, we retrieve only the frames containing landmarks. Smaller landmarks will have considerably less number of straight lines when compared to larger landmarks. Therefore we found out the local maximum frames of the edge map to extract larger salient landmarks and discard other unwanted objects and smaller landmarks.

Using the above methods we could identify good landmarks along the route. These landmarks are indexed in a chronological order so that they can guide the navigation from the starting point to the destination point using a low cost matching technique.

2.2 Landmark Matching

Speeded Up Robust Features (SURF). We did landmark matching using SURF feature matching. SURFs are in this work used for describing keypoints in landmark based navigation. It is a scale and in-plane rotation invariant detector and descriptor with strong repeatability and robustness. The

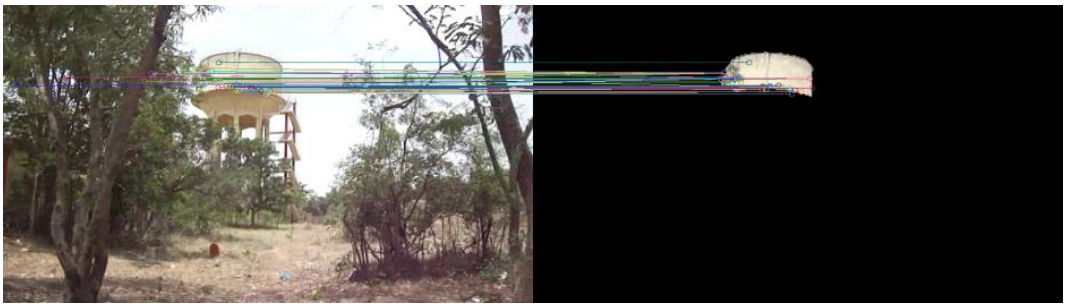
common application of SURF works as follows:

1. Interest points (such as corners, blobs and T-junctions) are selected (detection),
2. The neighbourhood of every interest point is represented by a feature vector (description),
3. Keypoints with closest Euclidean distances (above a given threshold) are matched.

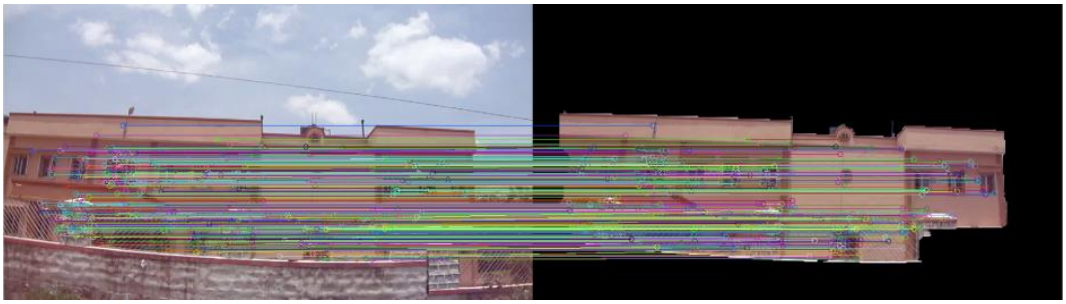
Concerning the photometric deformations, a simple linear model with the bias (offset) and contrast (scale factor) change is assumed. Neither the detector nor the descriptor use colour information. The detection is based on the Hessian-matrix approximation using integral images allowing fast computation of box type convolution filters, which drastically reduce the computation time. As the SURF is used for the image matching based on the Euclidean distance between the descriptor vectors, the dimension of the descriptor has a direct impact on the computational time. Here lies the main benefit of using SURF against still widely used SIFT (Scale Invariant Feature Transform) [20]. SIFT is a predecessor of SURF based on the local oriented gradients around the point of interest, which uses 128-dimensional vector meanwhile the SURF, based on the first order Haar wavelet [30] responses in horizontal and vertical direction, provides descriptor of dimension 64. In [22] is the comparison that shows similar results in precision for nearest neighbour matching for both SURF and SIFT. Nearest neighbour is an optimization problem for finding closest points in n-dimensional spaces. In many cases, the distance is measured by Euclidean or Manhattan distance.

As mentioned above, SURF features of the identified landmarks are detected and matched with the video frames for validating the navigation. Number of corresponding points determines matching between the candidate landmark and the landmark from the database. The corresponding points are more for a correct match(see figure 2.11b). The amount of correspondence to qualify as a correct match also depends on the area of the landmark. Larger landmarks should have larger correspondence and it is acceptable if there is less correspondence for smaller landmarks(see figure 2.11a). If the number of corresponding points is very less compared to the area of landmark(see figure 2.11c), then we consider such a scenario as landmark mismatch.

The matching process validates the rightness of the landmarks and assists in providing a better navigational guidance to the user. Though our navigational system automatically provides good landmarks to guide drivers, it does not provide decision making capability that works like a human brain. The user would like to make decisions manually by considering the information provided by our landmark-based navigation system. In the next chapter, we have provided a framework based on hand gestures to intuitively interact with the landmark-based navigation system.



(a) Good match for landmark with small area



(b) Good match for landmark with large area

(c) Landmark mismatch

Figure 2.11: Landmark matching using SURFs

Chapter 3

Hand Gesture Recognition system

3.1 Introduction

Computer technologies have grown tremendously over the past decade. As the computing, communication and display technologies progress even further, existing HCI techniques are becoming a bottleneck in the effective utilization of the available information ow. Typical HCI has been the norm all this while and people are unreasonably curious on how things can be done to change the nature of HCI. The most common mode of HCI is relying on simple mechanical devices, i.e. keyboards and mice. These devices have grown to be familiar but are less natural and intuitive in interacting with computers. Gesture enabled HCI transcends barriers and limitations by bringing the user one step closer to actual one to one interactivity with the computer. There have been much active research towards novel devices and techniques that allow gesture enabled HCI in recent years. The term Gesture recognition collectively refers to the whole process of tracking human gestures to their representation and conversion to semantically meaningful commands. Research in hand gesture recognition aims to design and development of such systems than can identify explicit human hand gestures as input and process these gesture representations for device control through mapping of commands as output. Creation and implementation of such efficient and accurate hand gesture recognition systems are aided through two major types of enabling technologies for human computer interaction namely hardware-based and vision-based approaches. Hardware-based approach requires user to wear bulky devices, hindering ease and naturalness of interacting with the computer. Although the hardware-based approach provides high accuracy, it is not practical in users everyday life. This has led to active research on more natural HCI technique, which is computer vision-based. This approach uses cameras and computer vision techniques to interpret gestures. Research on vision-based HCI has enabled many new possibilities and interesting applications. Some of the most popular examples are tabletop, visual touchpad, TV remote control, augmented reality and mobile augmented reality. Vision-based HCI can be further categorized into marker-based and marker-less approach. Several studies utilize color markers or gloves for real time hand tracking and gesture recognition. This approach is easier to implement and has better accuracy, but it is less natural and not intuitive.

3.2 Related work

Most of the research studies on marker-less hand gesture recognition focused on different techniques such as Haar-like features [23], Convexity defects [24], K Curvature [25], Bag-of-features [26], Template Matching [27], Circular Hough Transform [28], Particle Filtering [29], and Hidden-Markov Model. Several researchers use Haar-like features, which requires high computing power. The classifier preparation stage also consume a lot of time. Some studies use K-curvature to find peaks and valleys along a contour, and then classify these as fingertips. However, it is also CPU intensive because all points along the contour perimeter must be calculated. Besides, they did not solve the problem of differentiating between human face and hand regions because they assumed that only hand regions are visible to the camera. The work presented in this chapter mainly focuses on the following.

a) Implementation of a low cost marker-less vision based hand gesture recognition system which is flexible, intuitive and able to interface with other applications via different methods.

b) Employing simple but robust tracking methods.

c) Differentiating head and face regions.

3.3 System design

The overall system consists of three modules: Camera module, Detection module and Interface module. They are summarized as follows:

i. Camera module: This module is responsible for data acquisition. It captures a real time video frames from a regular webcams, and then processes this output with different image processing techniques. The output of this module is a smoothed binary image with clean contours suitable for hand gesture detection which is passed to the detection module frame by frame.

ii. Detection module: This module is responsible for detection and tracking of hand and fingers using our proposed method. Finite State machine and Camshift are applied to improve the accuracy and stability of tracking. The output of this module is hand and finger locations, orientation and motion in 2D space.

iii. Interface module: This module is responsible for translating the detected hand gesture into functional inputs and interfacing with the vehicle navigation system.

The hardware requirements are minimal; the system consists of a detector (low-cost USB webcam or depth camera), a computing apparatus (desktop), and a video display (monitor or projector).

3.3.1 Camera module

In this module, image frames are being retrieved from a USB webcam, and then processed through several steps using image processing techniques. Then, the pre-processed image frame is passed to the next module for detection. Each processing step in this module is discussed in detail in the next sub sections.

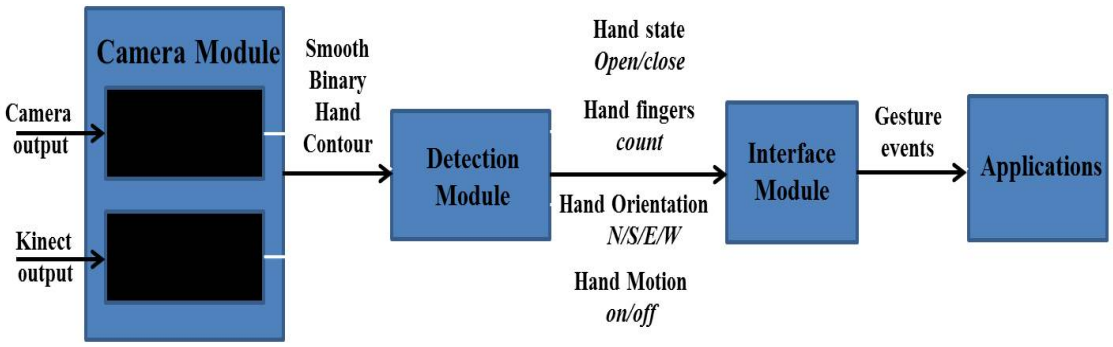


Figure 3.1: System design

a. Data acquisition

Our main motivation is to design a system that is able to utilize low-cost USB web-cams, which is readily available with most users. However, as low-cost depth cameras are becoming more common nowadays, we can also include the support for depth cameras in our system. The depth camera is slightly better in terms of hand segmentation from the background, but it suffers from lower resolution and frame rates. It results in less smoother HCI. It is also significantly more expensive than a USB web-cam. At this step, image frames(see figure 3.2) are being retrieved from the camera at 3060 frames per second (fps), depending on camera type. It is then passed to the next step for background subtraction. For the case of depth camera, only the depth segmentation step is required.

Figure 3.2: Captured image

b. Background subtraction

Background subtraction is performed to effectively segment the user from the background(see figure 3.3). Typical methods use a static background image and calculate the absolute differ-

ence between the current frame and the background image. Usually RGB color space is used while some studies propose HSV or YCrCb color space as a more efficient. Nonetheless, all these color spaces still possess limitations because color leakage will occur if the foreground object contains colors similar to the background. In our approach, we calculate the absolute difference (IABS) foreground and background images in RGB color space and converts it into grayscale image (IG). Then we convert IG into a binary image BW by replacing all pixels in the grayscale image (IG) with luminance greater than threshold level (L) with the value 1 (white) and all other pixels with the value 0 (black). The threshold level (L) is calculated using Otsu's method. Otsu's method is used to automatically perform clustering-based image thresholding i.e. the reduction of a graylevel image to a binary image. The algorithm assumes that the image to be thresholded contains two classes of pixels or bi-modal histogram which are foreground and background and then calculates the optimum threshold separating those two classes so that their combined spread (intra-class variance) is minimal.

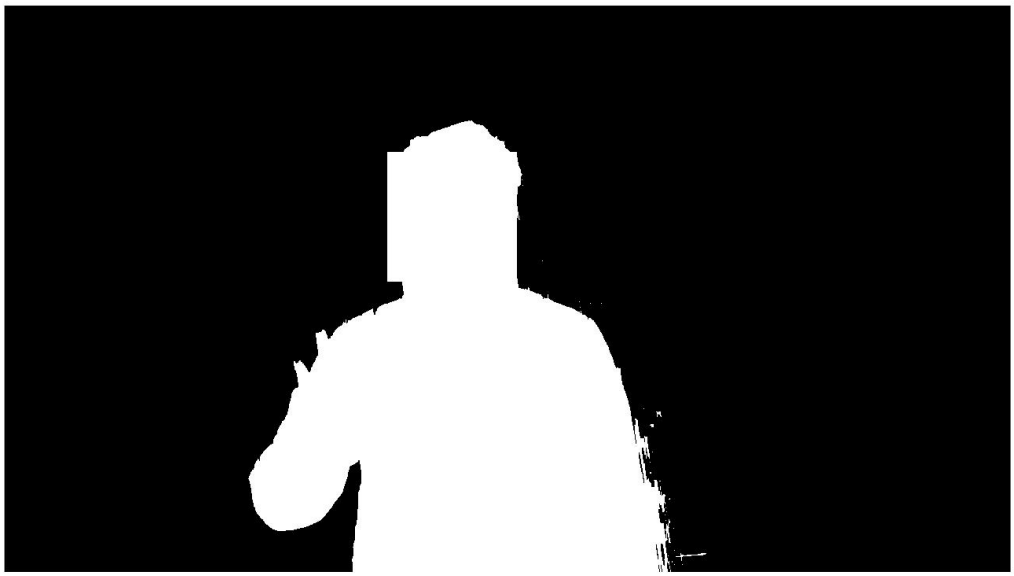


Figure 3.3: Foreground mask

c. Face removal

Our hand region extraction method is based on skin color extraction; therefore, both hand and face regions will be extracted. In some cases, it is hard to differentiate between a closed fist and the face. Therefore, we utilize Haar-like features [23] by Viola Jones [30] to detect the face region and then remove it by simply making the face pixels zero valued. Hence, face region will not be extracted during the next step (skin color extraction (see figure 3.4)). Haar-like features is efficient and it can detect human faces with high accuracy and performance. Preparing a good Haar classifier is a time consuming task, so we use the well-trained classifier provided in Matlab Computer vision toolbox.

d. Skin segmentation

Studies show that YCrCb color ranges are best for representing the skin color region. It also

provides good coverage for different human races. The basic value chosen in our implementation is based on the value suggested by D. Chai [31] as shown in figure. 3.4). The value is used as a threshold to perform skin color extraction in YCrCb color space and it is very efficient in covering a wide range of skin colors. However, it causes any object that contains a color similar to skin such as orange, pink and brown to be falsely extracted. This effect is significantly reduced as we have done background subtraction. We can also modify the value by setting a narrow default range so that it will efficiently extract only the skin region but not others objects.

Figure 3.4: Skin pixel mask

e. Morphology operations and Extracting the hand region
   In order to remove noise efficiently, we apply a morphology Opening operator (Erosion followed by Dilation) in several stages; during background subtraction and after skin extraction. After skin segmentation we get all the skin regions but as per the camera position, the hand and the face are the only large skin regions in the image. Therefore we can extract the hand region(see figure 3.5) by retaining the largest skin region and discarding the others.

If we had the depth information from the depth camera (Kinect), we can easily remove the background and keep only the hand contours as in section e by applying a fixed depth threshold range from the nearest object to the camera, because in this scenario the hand has the least depth from the camera. Anything beyond this depth range will be discarded and not taken into consideration when performing hand detection in the next module. Hence, steps from section b to e could be omitted.

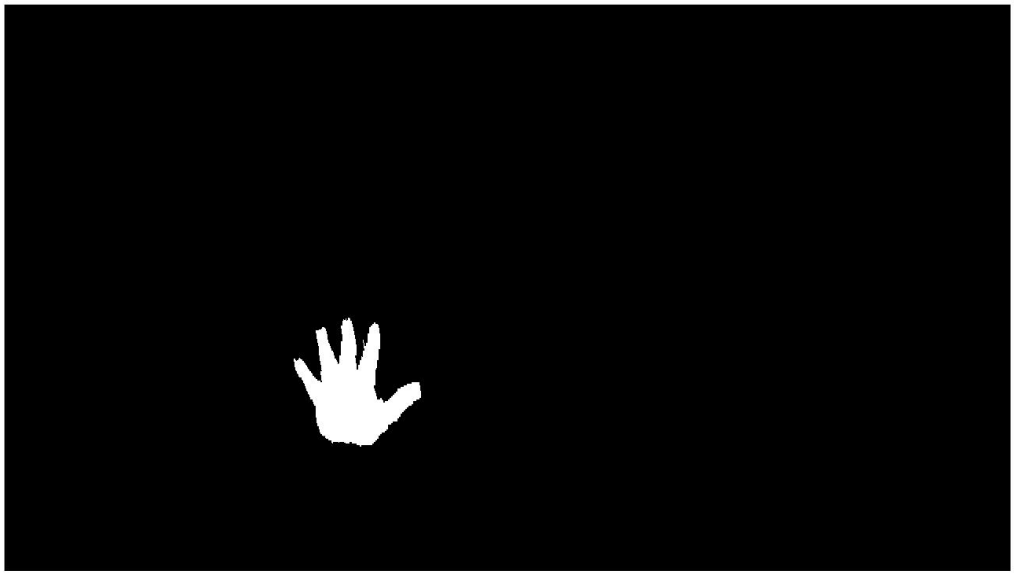


Figure 3.5: Hand contour

3.3.2 Detection module

The output from the camera module is a binary image with smoothed and polished hand contours. In this module (Fig.6), more information will be extracted, such as hand locations, hand open/close state, finger counts, fingers locations, fingers orientation, and hand movement.

a. Distance transform
   Now the output of the camera module is processed through Distance transform method. The distance transform method gives the distance (Euclidean distance) of each pixel from the nearest boundary pixel. The distance from the boundary to a pixel in the hand region increases as the pixel is away from the boundary(see figure 3.6). Using this distance value, the centroid of the palm region can be calculated.

0	0	0	0	0	0	0	0
0	1	1	1	1	1	1	0
0	1	1	1	1	1	1	0
0	1	1	1	1	1	1	0
0	1	1	1	1	1	1	0
0	1	1	1	1	1	1	0
0	0	0	0	0	0	0	0

→

0	0	0	0	0	0	0	0
0	1	1	1	1	1	1	0
0	1	2	2	2	2	1	0
0	1	2	3	3	2	1	0
0	1	2	2	2	2	1	0
0	1	1	1	1	1	1	0
0	0	0	0	0	0	0	0

Figure 3.6: Distance transform of a rectangular contour

b. Palm center and radius determination
   The image I_D of the hand after applying distance transform is shown in figure 3.7. The white color in the center is intense and the color fades as the distance from the center increases.

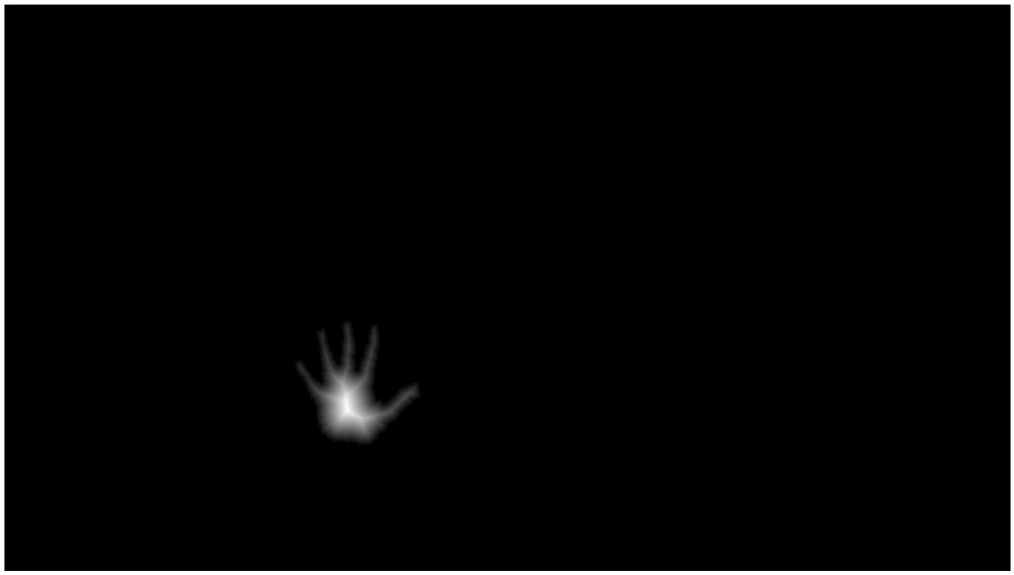


Figure 3.7: Distance transform of the hand

From this it is evident that the pixels near the boundary have lower values for distance and the pixels away from the boundary have higher values for distance. This middle region which has the highest value for distance is considered as the centroid of the palm. The width of the hand region will be approximately twice the distance from centroid to the nearest boundary pixel (say 2d) as shown in Figure 3.8 . Therefore the radius of the palm is d.

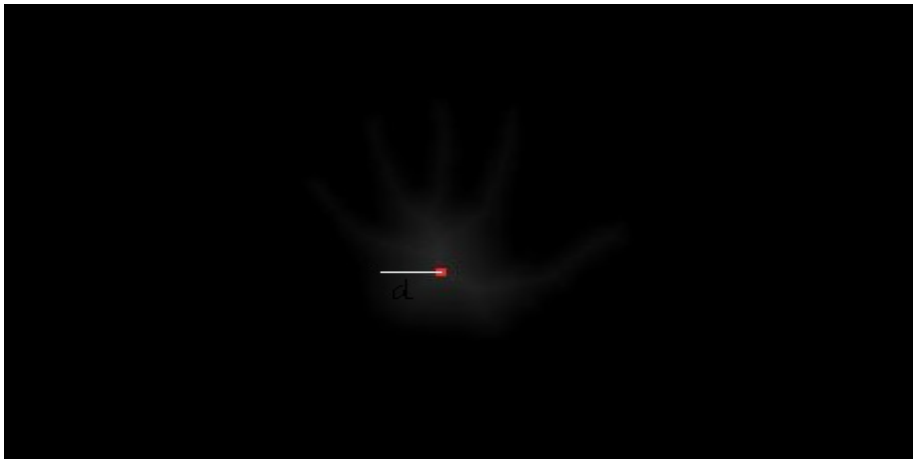


Figure 3.8: Palm center and radius

c. Fingers counting

The width of each finger is approximately one fourth of the width of the hand (i.e. $\frac{1}{4}^{th}$ of $2d$). Now a suitable structuring element S (disc) that can erode the fingers completely is chosen and erosion is performed on the segmented hand region.

After erosion only a part of the palm region RP1 is left behind and the finger regions are completely eroded. Further the palm region which remains after erosion RP1 is dilated using the same structuring element and this give the region RP2 which is larger than the dilated palm region.

The dilated palm region RP2 is subtracted from the original binary image IB to give the finger regions FR alone. Now the total number of components is the number of fingers. They are represented by drawing a line along the major axis of the segmented finger regions as shown in figure 3.9. The number of lines drawn is equal to number of active fingers.

(a) 5 fingers (b) 2 fingers

Figure 3.9: Finger regions

d. Hand orientation

The orientation of the hand is based on the relative position of the center of mass of the fingers to the centroid of the palm(see figure 3.10).

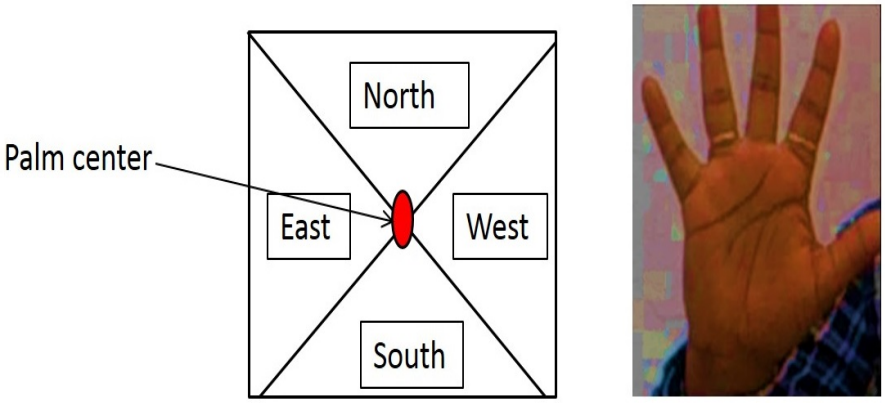


Figure 3.10: Hand orientation

e. Hand motion

The motion of the hand is determined by the variation in the distance between the palm centers. If this distance is larger then a certain threshold value, then the hand is considered to be in motion.

f. Finger orientation
We calculate the pairwise angles between the all fingers and determine the finger oreintaitons(see figure 3.11).

(a) {Angles: 20, 18, 25, 89}

(b) {Angles: 25}

Figure 3.11: Finger orientation

3.3.3 Interface module

Interface module is responsible for composing hand gestures based on the features from the detection module and translating them into meaningful actions appropriate to the application. As several degrees of freedom with hand gestures are possible, a natural choice would be to use a multilayer decision classifier. Such a classifier takes decisions at every layer of the classifier based on the most basic hand features (such as hand movement) at layer one to complicated hand features at deeper layers.

Once the features of the hand(F_n) are identified in the detection module, the system is ready for gesture recognition. First layer determines the motion of the hand. Absence of motion indicates no gesture. When the hand is in motion, second layer checks the orientation of the hand. Orientation is decided based on angle made by the line joining centroid of the palm with center of mass of the middle finger, with the horizontal axis. Based on the angle, classifier decides orientation to be north (N), south (S), east (E) and west (W). The third layer determine the number of fingers based on the number of connected components. After the application of distance transform followed by certain morphological operations on the hand contour, only finger segments will remain in the binary image. Counting such segments gives number of fingers. Final layer calculates the direction vectors of all fingers and all pairwise angles between extended fingers. In future, we plan to include additional features such as finger name, etc. Utilizing such a layered classifier makes sense because the number of possibilities at upper layers is less (two in this case: motion absent and motion present) and increase with layer number. Our four layer configuration gives a size four feature vector which is

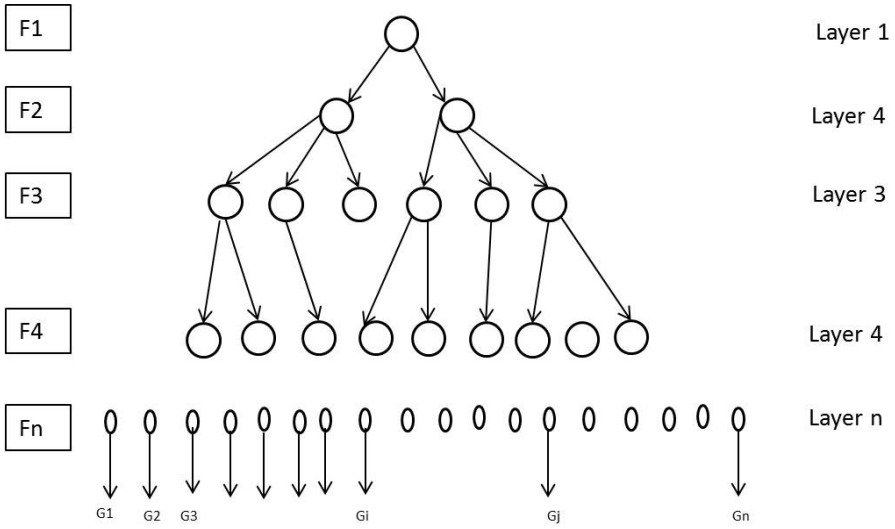


Figure 3.12: Decision-based classifier

then mapped to some action. For example, classifier output can be {Motion: present, Orientation: north, Number of fingers: 3, Angles: 20, 12}.

Figure 3.13: Sign-Language conversion,Gesture: "Hi"

In order to demonstrate our gesture recognition system, we have developed a simple Sign to language translator application which would be useful for people having hearing and speech impairment. The deaf-dumb have their own manual-visual language known as sign language. Sign language is a non-verbal form of intercourse which is found amongst deaf communities in world. The languages do not have a common origin and hence difficult to interpret. Hence, It is very difficult for a normal person to understand the sign language of the deaf and dumb community and interact with them. Our application recognizes the sign language gestures and converts them into appropriate text by which the normal person can easily interact with people belonging to deaf and dumb community. Currently we have only included four gestures which would be helpful to convey greetings through signs. The word "Hi" is modeled by assigning features as {Motion: present, Orientation: north, Number of fingers: 5, Angles: 20, 18, 25, 89}. The word "How are" is modeled by assigning features

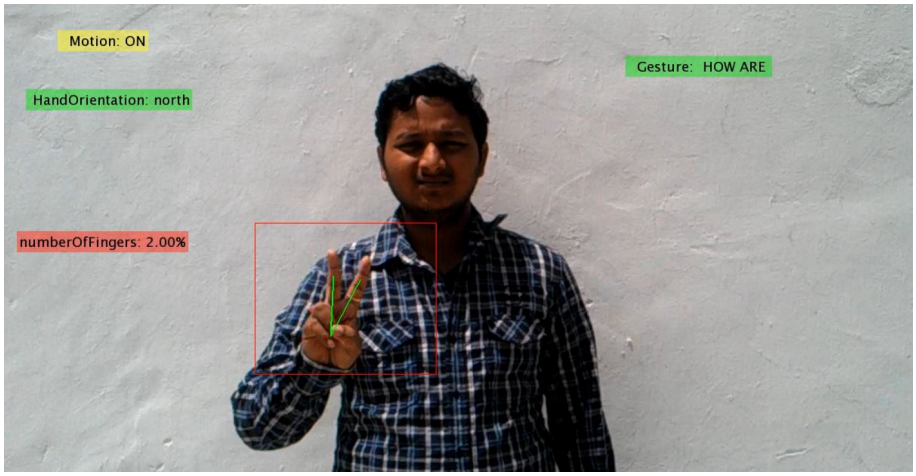


Figure 3.14: Sign-Language conversion,Gesture: "How are"

as {Motion: present, Orientation: north, Number of fingers: 2, Angles: 25}. The word "You" is modeled by assigning features as {Motion: present, Orientation: north, Number of fingers: 1, Angles: none}. The word "Bye" is modeled by assigning features as {Motion: present, Orientation: north, Number of fingers: 4, Angles: 21, 17, 24}.

Figure 3.15: Sign-Language conversion,Gesture: "You"

Our hand gesture recognition system can interface with many such applications. We have also integrated it to the proposed landmark-based navigation system where a manual intervention of the user is required to validate the identified landmarks for efficient navigation.

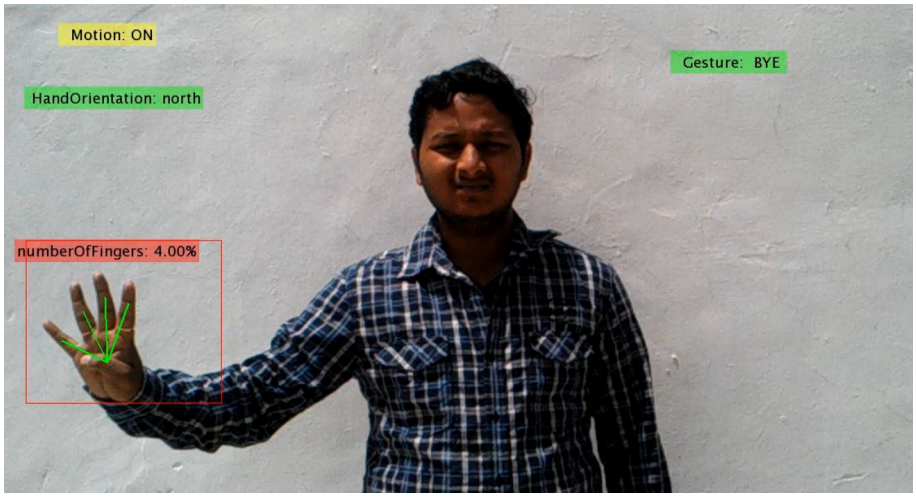


Figure 3.16: Sign-Language conversion,Gesture: "Bye"

Chapter 4

Results

By developing a landmark recognition system that combines salient region detection, segmentation, and edge based non-landmark object removal techniques, we are able to recognize and localize landmarks from images and videos. In this chapter, we illustrate our landmark-based navigation system in next generation vehicles using experimental results on real world data. To the best of our knowledge there is no standard database available to validate the proposed method. Hence we collected our own data for validation. Sony cybershot, 16 MP digital handycam with 60 fps was utilized to capture videos of the navigation path.We initially took videos at a vehicle speed of above 60 kmph, but encountered motion blur. Therefore for initial evaluation of our system we captured various videos at a vehicle speed of 30 kmph.

(a) Landmark building (b) Segmented region of landmark

Figure 4.1: Saliency based landmark (building) segmentation

First, saliency based landmark segmentation is presented. Figure 4.1 and 4.2 show examples of such a segmentation. The building and tank are salient as compared to their background. Interestingly, the tank is far from the camera and yet gets segmented in the most of the frames. It is not identified in a few frames where most of the tank is occluded by trees.

Further, once such landmarks are identified, the next task is to match them with the candidate landmarks. In next few figures, we show results for such matching. The reference landmark is matched with candidate landmark using Speeded Up Robust Features (SURF). A large number of

(a) Landmark tank (b) Segmented region of tank

Figure 4.2: Saliency based landmark (tank) segmentation

matched would indicate correct match. However, this also depends on the area of the landmark in the image. For example, one observes large number of matches in building landmark, where the area of the building is large. For smaller landmarks such as tank, the number of matches is small. Thus, we consider area normalized number of matches.

Such a procedure sometimes might yield incorrect results in few cases. For example, two altogether different buildings may have similar windows and color. In such cases, we choose manual intervention. A landmark match proposed by the system is either accepted or rejected by user by showing some gesture. Here, we use 'raised thumb' to approve the match, 'thumb upside down' to disapprove the match and 'horizontal thumb' for good landmark however bad match.

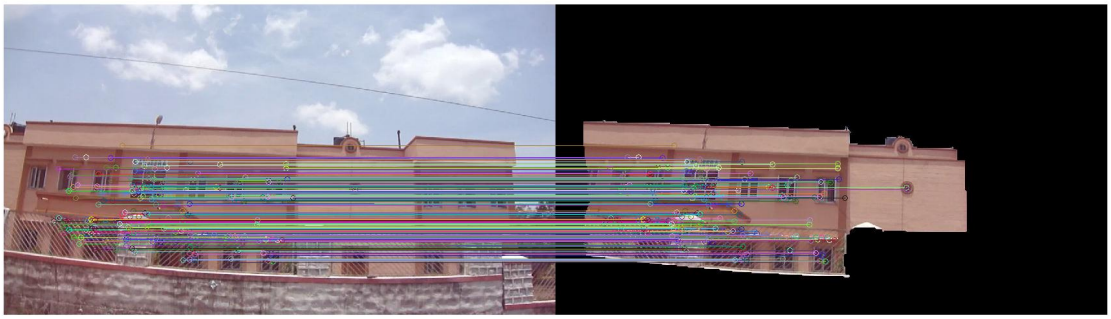


Figure 4.3: Good match

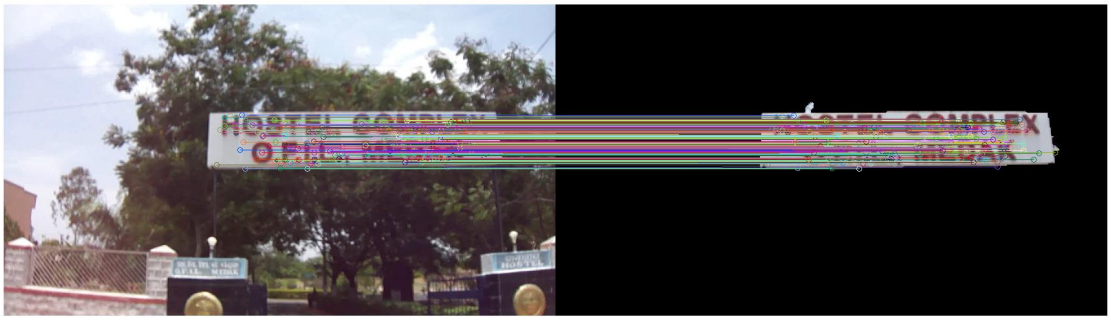


Figure 4.4: Good match

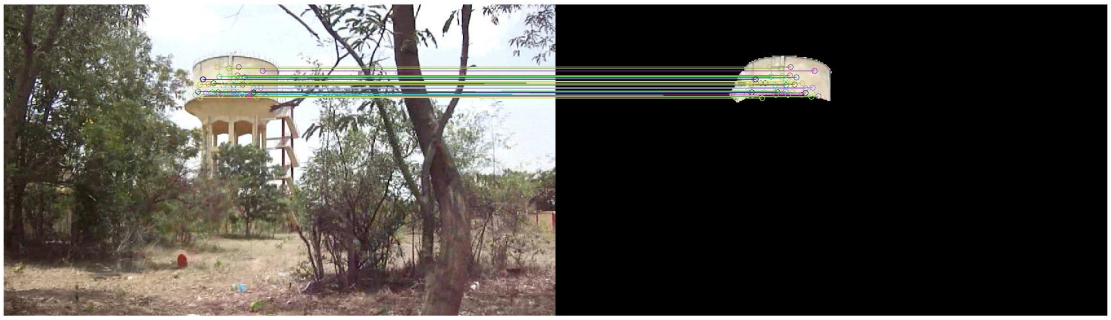


Figure 4.5: Good match

Figure 4.6: Landmark mismatch

(a) Good match; action–search next landmark

(b) False match; action–keep searching

(c) Good match but not a good landmark; action–keep searching

Figure 4.7: Gesture inputs for landmark matching

Chapter 5

Conclusions and Future Work

The following are some of the contributions that have been made in the course of my exploration of this thesis.

5.1 Summary of contributions

- **Saliency-based landmark object segmentation:**

  In the Chapter of Landmark-based navigation system, we have presented Saliency-based landmark object segmentation based on Global Contrast based Salient Region Detection by Cheng et al[16] . Given a video frames along the navigation path, our method segments a salient landmark in the frame. Here, the image is divided into many regions using Efficient graph-based image segmentation by Felzenszwalb et al[17]. Grabcut technique[18] has been employed to segment the most salient region after assigning saliency scores for each region in the frame.

- **Edge-based landmark identification:**

  Saliency based object segmentation also segmented many unwanted objects like trees, sky, vehicles, etc. In the Chapter of Landmark-based navigation system, we also proposed an edge-based technique to discard the frames containing unwanted objects and retain only large salient landmarks. Here, we exploited the fact that the edges of landmark buildings are mostly regular structures such as rectangles and straight lines.

- **Human-navigation system interaction:**

  A navigational system guides the user for better navigation, but cannot provide a decision-making capability as robust as human mind. Therefore, there is a need for humans to interact with the navigation systems for safe and efficient navigation to the final destination. Interaction through gesture is more intuitive and natural when compared to orthodox forms of interaction. In the Chapter of Hand gesture recognition system, we provided a framework for intuitive interaction with the navigation system using natural hand gestures.

5.2 Future Work

- **For salient object segmentation in images and videos, future directions may include:**

  To segment the image in regions using a region growing technique, before calculating the region contrast saliency.

- **For discarding frames with unwanted objects, future directions may include:**

  To use dictionary-based learning and structure from motion rather than the proposed edge-based technique.

- **For better gesture recognition, future directions may include:**

  To develop a random forest learning based gesture classifier instead of simple decision based classifier. This would improve the quality of gesture recognition.

- **For better navigation, future directions may include:**

  To integrate the detected landmark with a digital navigational map to know the exact position of the landmarks.

5.3 Conclusion

In the last decade, car navigation systems and tools have evolved considerably as briefly shown in Figure 5.1. Many car navigation systems have become available. These systems offer the promise of easily-accessible and friendly multimodal user interface, but the existence of such diverse navigation tools raises the question of what is the better way to provide route guidance and navigation information to the drivers?

Figure 5.1: Timeline for navigation systems and tools

Simple portable navigation devices are capturing the world car navigation market, accounting for about 80% of the 45 million devices worldwide. This situation is however changing. As the 4G mobile telephony network expands, smart-phone companies such as Samsung, Apple and other similar mobile terminals are beginning to offer the same or sometimes better navigation functions as these simple portable navigation devices. The change in the car navigation market is affecting not only portable devices, but installed models as well. As a result, the whole car navigation systems market is becoming a structure combination of: mobile phones, portable devices, and factory-installed navigation systems.

Due to the rapid advances in the automobile industry, there is a need for robust, low cost navigational system. In this thesis we have tried to give a solution for landmark-based navigation in next generation vehicles. We started by presenting a landmark-based navigation system which identifies reference landmarks along the destination path and provide a navigational guidance by matching the reference landmarks and validating the path. We presented saliency-based landmark object segmentation method for landmark identification. Unwanted objects obtained during landmark segmentation have been removed by processing the frames through an edge based landmark identification method. The identifies landmarks are provided to the user to guide his navigation by matching them along his path using SURF feature matching. The proposed system provides a semi-automatic navigational guidance where the user has to manually validate the matching and rightness of the landmarks. For this, we have presented an interactive platform based on hand gestures to manually interact with the navigational system.

The gesture recognition system consists of three modules: 1) Camera module, 2) Decision module and 3) Interface module. The camera module captures the video frames of the user, processes them through various image processing techniques and outputs the binary image of hand contour. The decision module is responsible for identifying certain features of the hand such as hand motion, finger count, hand orientation, finger orientation and finger name. The interface module composes hand gestures based on these features by constructing a multilayer decision-based classifier. We have also illustrated sign to language conversion application using the hand gesture recognition system which would be useful to hearing and speech impaired people to communicate with the normal world. Several Hand gesture recognition techniques have also been explored. The proposed gesture recognition technique provide a platform to interact with many applications requiring gestures to assign meaningful actions.

Within the course of this thesis research, some obstacles have been found in the evaluation of proposed methods. The lack of standard datasets for evaluation of landmark-based navigation hinders development of the field. The author of this dissertation with collaborators has worked hard to collect real-world data. We believe that in the end, the most important part of the work done for this thesis is to offer some technical possibilities of our dissertation question that is not fully answered yet, what is the better way to provide route guidance and navigation information to the drivers? We have tried to answer this question based on the resources available to us. However, fully answering the question is beyond the scope of one single thesis and requires tremendous amount of research technology advances. We hope that we have succeeded in presenting some useful techniques for landmark-based vehicle navigation and an interesting new way of interacting with the vehicle navigation system. We look forward to contributing further to improve driving safety, efficiency and overall experience.

References

[1] L. A. Streeter ,D. Vitello, "A profile of drivers map-reading abilities", *Human factors*, 1986.

[2] A. J. May, T. Ross, "Presence and quality of navigational landmarks: Effect on driver performance and implications for design", *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 2006.

[3] G. E. Burnett, "Turn right at the King's Head: Drivers' requirements for route guidance information", *PhD thesis, Loughborough University, UK*, 1998.

[4] K. Lynch, "The Image of the City", *MIT Press*, 1960.

[5] S. Kaplan, "Adaption, structure and knowledge. G. Moore & R. Golledge (Eds.) Environmental knowing: Theories, research and methods, 1976.

[6] Q. Iqbal and J. K. Aggarwal. Applying perceptual grouping to content-based image retrieval: Building images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1999.

[7] D. Robertson and R. Cipolla. An image-based system for urban navigation. In Proceedings of the British Machine Vision Conference (BMVC), 2004

[8] T. T. H. Shao, T. Svoboda, and L. V. Gool. Hpat indexing for fast object/scene recognition based on local appearance. In Computer lecture notes on Image and video retrieval, 2003.

[9] A. Torralba, K.P. Murphy, W.T. Freeman, and M. Rubin. Context-based vision system for place and object recognition. In Proceedings of the IEEE Intl. Conference on Computer Vision (ICCV), 2003.

[10] Y. Li and L. G. Shapiro. Consistent line clusters for building recognition in CBIR. In Proceedings of Intl. Conference on Pattern Recognition (ICPR), 2002.

[11] R. Achanta, F. Estrada, P. Wils, and S. S usstrunk. Salient region detection and segmentation. In ICVS, pages 6675. Springer, 2008.

[12] R. Achanta, S. Hemami, F. Estrada, and S. S usstrunk. Frequency-tuned salient region detection. In CVPR, pages 15971604, 2009.

[13] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu. Sketch2photo: Internet image montage. ACM Trans. Graph., 28(5):124:110, 2009.

[14] A. M. Triesman and G. Gelade. A feature-integration theory of attention. Cognitive Psychology, 12(1):97136, 1980.

[15] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurbiology, 4:219227, 1985.

[16] Ming-Ming Cheng, Guo-Xin Zhang, Niloy J. Mitra, Xiaolei Huang, and Shi-Min Hu, "Global Contrast based Salient Region Detection," *IEEE CVPR*, pp. 409–416,2011.

[17] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. IJCV, 59(2):167181, 2004.

[18] C. Rother, V. Kolmogorov, and A. Blake. Grabcut Interactive foreground extraction using iterated graph cuts. ACM Trans. Graph., 23(3):309314, 2004.

[19] J. Canny, A Computational Approach To Edge Detection, IEEE Trans. Pattern Analysis and Machine Intelligence, 8(6):679698, 1986.

[20] D. G. Lowe, "Object recognition from local scale-invariant features," *IEEE international conference on Computer vision (ICCV)*, vol. 2, pp. 1150–1157, Kerkyra, Greece, 1999.

[21] Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features.

[22] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). Comput. Vis. Image Underst., 110(3):346359, June 2008.

[23] Barczak ALC, Dadgostar F (2005) Real-time hand tracking using a set of cooperative classifiers based on Haar-like features.

[24] Manresa C, Varona J, Mas R, Perales F (2005) Hand tracking and gesture recognition for humancomputer interaction.

[25] Han SI, Mi JY, Kwon JH, Yang HK, Lee BG (2008) Vision based hand tracking for interaction.

[26] Dardas NH, Alhaj M (2011) Hand gesture interaction with a 3D virtual environment.

[27] Hasan MM, Mishra PK (2012) Real time fingers and palm locating using dynamic circle templates.

[28] Burns A-M, Mazzarino B (2006) Finger tracking methods using eyesweb. Gesture in humancomputer interaction and simulation.

[29] Bretzner, L, Laptev I, Lindeberg T (2002) Hand gesture recognition using multi-scale colour features, hierarchical models and particle filtering.

[30] Viola P, Jones MJ (2004) Robust real-time face detection.

[31] Chai D, Ngan KN (1999) Face segmentation using skin-color map in videophone applications.